Some principles for regression diagnostics and influence analysis: Comments on "Developments in Linear Regression Methodology: 1959-1982"

Sanford Weisberg

University of Minnesota

Technical Report No. 416

February 1983

University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota 55108

Ron Hocking's own work on regression has played such an important role in the development of regression methods that it is very fitting for him to have written this review of the last quarter century of advances. His 1976 review paper on selection methods in Biometrics got me interested in that problem and possibly in regression in general, and that paper, like the current one, is exemplary. Both survey an area, but still leave many questions unanswered.

In fitting linear regression models, we make many assumptions, such as linearity, constant variance and perhaps normality. We assume that relevant variables are measured, and that these do not need to be transformed. As Hocking has pointed out, the precomputer approach of taking the assumptions as given and correct is no longer accepted statistical practice. Methods for criticism of assumptions and of influence analysis have now become standard, as clearly indicated by the proportion of Hocking's review that is dedicated to such methods. Hocking does note, however, that the array of such techniques that are available to the analyst is large, so the choice of appropriate and useful measures is not very clear. The confusion has several sources. The whole methodology of regression criticism has developed very quickly. Before the last decade, the most common tools for criticism were plots of residuals against various quantities such as fitted values, and probability plots. Each of these was intended to serve a number of purposes, providing information on outliers, linearity, heteroscedasticity, the need to transform, and perhaps some notion of influence, depending on the pattern in the plot. The recently developed or rediscovered methods for criticism, on the other hand seem to address specific issues, and each of these methods may require computation of statistics useful for that one

method only. At the same time, methods have been developed that are probably not generally useful, but many regression analysts are not sufficiently knowledgeable to tell the good ones from the not so good ones. My purpose in these remarks is to present some guidelines for developing, and using, methods for regression criticism. I will separately address what I call diagnostics (model criticism) and influence analysis (data criticism).

Before proceeding, it may be well to point out that not everyone agrees with the importance of these methods. Some think that they do little more than allow analysts to make quick but superficial decisions concerning data. I obviously do not agree with this view, and have found that intelligent application of these methods can be very useful in practice. In any case, the discussion following Atkinson (1982), is illuminating on this issue.

1. Diagnostics

Regression diagnostics, as defined here, are statistics designed to help an analyst decide if assumptions made in fitting a model are tenable. Good diagnostics will both suggest a problem, and a possible solution to the problem, such as a transformation. Following Box (1980), these methods are done conditionally given the fitted model, so any principles that are to be developed for them are likely to be the same for both a Bayesian and a frequentist; prior information helps determine the fitted model but possibly not its criticism. Necessarily, the methods of criticism will be functions of residuals and other related quantities, since these contain the information in the data not modelled by the fitted model. Diagnostic methods consist of combining the residuals in various ways to produce tests or graphs that bear on various assumptions.

Cox (1977, Sec. 1.6) presents a typology of methods of assessment of model adequacy. The first of these might be called ad hoc methods, in which plausible statistics are computed whose distribution is known, at least approximately, under a correct model. A graphical example of an ad hoc method is the now standard plot of ordinary residuals against fitted values. Under a correctly specified linear model, such a plot will show no pattern, while it is hoped that if the model is somehow incorrect, this plot will be systemmatic. For example, nonconstant variance may lead to a megaphone shape (Weisberg, 1980a, Chapter 6). But even this simple example shows the limitation of ad hoc methods. To understand this plot, we must know the distribution of the statistic, here a plot, when the model is not correct. It is not hard to specify alternatives for which the plot above would miss nonconstant variance, or miss any other specific model failure. This suggests the first principle for diagnostic methods:

D1. The behavior of a diagnostic procedure must be known, at least approximately, both under the correct model and under the model with one particular assumption modified.

Under D1, it is unlikely that a plot such as that of residuals versus fitted values will emerge as a recommended method for a specific purpose, since many different incorrect assumptions such as nonconstant variance or an outlying set of points may lead to the same graph. We need methods that are specific to one particular assumption, not general, omnibus methods.

This leads naturally to consideration of the second class of diagnostic methods, using what Cox called model expansion. The idea here is to consider fitting a different model with additional parameters, such that the new parameters will have a specific value when assumptions concerning the model are true. The best known example is obtained by writing an expanded linear model as

$$Y^{(\lambda)} = X\beta + \epsilon$$

where $Y^{(\lambda)}$ is some transformation of the data vector Y, such that $Y^{(\lambda_0)} = Y$ for some λ_0 and all Y. This is of course the basis of the Box and Cox (1964) method for determining a transformation toward normality or linearity, depending on the context. A test of $\lambda = \lambda_0$ provides a diagnostic test of the need to transform, and the estimate of λ provides the estimated transformation. Because of the parametric form of the alternative, it is reasonably clear what failure of the particular assumption studied would mean. This leads to our second principle:

D2. Useful diagnostics can often be derived by parameterizing the assumptions, thereby turning the problem of criticism, at least approximately, into one of parametric inference.

The Box and Cox procedure by itself does not supply the necessary tools for a diagnostic method, because estimation of λ can be nontrivial, especially if λ has dimension greater than 1. Diagnostics must be quick methods, not computationally intensive. One approach that works well is the so called Lagrange multiplier, or score tests (Aitchison and Silvey, 1960) or modifications thereof (Moran, 1970; Atkinson, 1973). The general method is as follows. Let $\theta^T = (\beta^T, \sigma^2)$, and suppose the log likelihood under the expanded model is $L(\theta,\lambda)$. Let $U = \partial L(\theta,\lambda)/\partial\lambda$ evaluated at $\lambda = \lambda_0$ be the score vector, and let

$$\left(-E\frac{\partial^2 L(\theta,\lambda)}{\partial(\theta^T,\lambda)^T\partial(\theta^T,\lambda)}\right)^{-1} = \begin{pmatrix} A(\theta,\lambda) & B(\theta,\lambda) \\ B^T(\theta,\lambda) & C(\theta,\lambda) \end{pmatrix}$$

be the inverse of the expected information at (θ,λ). Then a test of $\lambda = \lambda_0$ is computed as

$$S = \hat{U}^T C(\hat{\theta},\lambda_0)\hat{U}$$

where $C(\hat{\theta},\lambda_0)$ and $\hat{U}$ are evaluated at the mle $\hat{\theta}$ given $\lambda = \lambda_0$, or at some other consistent estimator of θ. An asymptotically equivalent test that may be of use in some problems would replace $C(\hat{\theta},\lambda_0)$ by the appropriate submatrix of the inverse of the observed information, evaluated at $\hat{\theta},\lambda_0$. In any of these cases, S has an asymptotic $\chi^2(\dim(\lambda))$ distribution.

The advantages of this approach are many. First, the parameters of the original model must be estimated only once under the assumed model, making computations simple even in some complex problems. Often, these tests have a very appealing form. Second, the tests are asymptotically equivalent to likelihood ratio tests, and therefore much is known about their asymptotic behavior. Even though small sample distributions of S is often intractable, large sample results are generally adequate for diagnostic purposes. Finally, the same approach can be applied in general regression problems, so new ideas are not required for other problems. This is summarized as the third principle for diagnostics.

D3. Diagnostic methods should not be computationally intensive. The class of score tests provide a rich class of diagnostic methods well suited to this purpose, and provides a standard against which other methods can be compared.

In Table 1 are listed references to various score tests that have been proposed for regression diagnostics. The score test for autocorrelation turns out to be equivalent to the Durbin-Watson statistic, and the score test for a single mean shift outlier in linear regression is the maximum Studentized residual. For this latter problem, the likelihood ratio test is the maximum deleted Studentized residual, which would be preferred because it has a known distribution. Atkinson's score test for transforming the response turns out to be equivalent to computing a regression with an added variable, and hence has a particularly nice form.

Table 1. Score tests

Assumption alternative	Score test reference	Equivalent plot
Autocorrelation	Box (1980)	Box (1980)
Transform response	Atkinson (1973,1982)	Box (1980) Atkinson (1982) Cook & Weisberg (1982)
Transform predictors	Box & Tidwell (1962)	Cook & Weisberg (1982)
Outliers	Cook & Weisberg (1982)	
Non-normality	Jarque & Bera (1980)	Probability plot; Atkinson (1981)
Heteroscedasticity	Breusch & Pagan (1980) Cook & Weisberg (1983)	Cook & Weisberg (1983)

We have argued that diagnostic methods should be more formal, so that their behavior can be known at least approximately. At the same time, they must be computationally simple to be useful in practice. The next characteristic that good diagnostics should possess is given as the fourth principle for diagnostics.

D4. Good diagnostics are graphical or have graphical equivalents.

At least in part, the need for graphical methods is the concern for outliers and influential cases: we do not want to diagnose the need to transform, for example, because of one unusual case. Graphs allow the analyst to make comprehensive checks of the data. Indeed, one may view the statistics as quantities that calibrate the plot, so the plot is the main diagnostic (see Cook and Weisberg, 1983). For some methods, the graphical equivalent is obvious. Since Atkinson's test is equivalent to adding a variable to a model, an added variable plot or a partial residual plot (see Cook and Weisberg, 1982, Sec. 2.3) are obvious choices. For non-normality, normal probability plots of residuals may have some benefit (but see Weisberg, 1980b and Atkinson, 1981). Since the score test for heteroscedasticity is equivalent to regressing squared residuals on independent variables, plots of squared Studentized residuals are suggested.

The fifth and final principle of diagnostics is

D5. Diagnostic procedures should suggest remedial action.

Atkinson's methods, for example, can be used not only to test the need to transform, but can also provide an estimate of the needed transformation. Unfortunately, most of the other score test diagnostics do not yet provide as much guidance, but further research may lead to useful procedures.

One important side effect of the development of separate procedures for each problem is that each method may require use of a <u>different</u> function of

the residuals. For example, in Hocking's notation, outlier testing uses the t_i^*, heteroscedasticity plotting is best done with the t_i^2, while heteroscedasticity testing uses the squres of the ordinary residuals. The analyst is charged with the problem of learning which set of residuals to use for each diagnostic.

The great advances in diagnostics, I believe, are in separating all of the assumptions that go into modelling, and designing separate, well defined methods for each. These will not take the art out of regression modelling, but they may tell us what colors will work reasonably well.

2. Influence

The general idea in influence analysis is to study the changes caused in the fitted model or other aspects of an analysis when the data are slightly perturbed. Whereas regression diagnostics are used to find problems with a model, influence analysis is done as if the model were correct; we study the robustness of the particular data set, in clmbination with a particular model, to the perturbations. This notion of robustness is closer to Box's original definition than the more popular current use with regard to robust estimators.

As with regression diagnostics, several guiding principles can be suggested for choosing influence measures.

I1. The perturbation scheme should be well defined.

The most popular perturbation scheme is based on deleting cases one at a time, or perhaps in small groups. We then study the behavior of estimators, or other quantities, computed without the deleted case. This perturbation scheme is very appealing on several grounds. First, we are led to statistics with values for each case, a desirable feature, since cases can then be identified

as influential. Second, this scheme is very easy to understand and to explain. Finally, it leads, in linear models, to very elegant results. Other schemes have met with only mixed success (e.g. Davies and Hutton 1975, Hodges and Moore 1972). For example, consider a perturbation scheme in which X is perturbed to X+E, where E is a matrix with small random elements. Usual perturbation theory from numerical analysis can be used to define potentially interesting measures of this sort of perturbation. Unfortunately, such methods require an estimate of the covariance matrix of E, and may lead to measures that are not invariant under linear transformation of the model. Invariance seems to me to be of paramount importance.

Once we agree on a perturbation scheme we must choose something to measure:

I2. Influence measures must refer to some specific aspect of the problem. They must measure something interesting.

If $\hat{\beta}$ is the estimator of β based on all the data, and $\hat{\beta}_{(i)}$ is the estimator without case i, then $(n-1)(\hat{\beta}_{(i)}-\hat{\beta})$ can be shown to be equal to the sample influence curve for β (Cook and Weisberg, 1982, Sec. 3.4.2). This difference measures a quantity that is clearly of interest: how much does the estimator change when we delete case i? It is not enough to find a measure that seems to behave correctly in examples; we need to know exactly what it is measuring.

There is no intrinsic reason why interest must center on β. Johnson and Geisser (1983) and Cook and Weisberg (1982) consider problems where prediction is of primary interest although the measures are very similar to those based on $\hat{\beta}$. Measures for changes in σ^2 have been proposed, but these are probably not used much in practice, since for most investigations estimation of σ^2 is not a primary concern. The Andrews and Pregibon (1978) measure attempts to give an omnibus measure of influence, and as such it does not

correspond to any specific aspect of an analysis. Its relevance to regression problems under principle I2 is therefore unclear.

The third guiding principle is:

I3. Influence measures should depend on the sample at hand.

There seem to be two divergent views on influence in general, depending on whether one chooses to condition on the (finite) sample, or study asymptotics or decision theoretic approaches. The latter view is exemplified by Huber (1983), and by using influence curves in place of sample influence curves. When this point of view is adopted, the natural measures of influence are the diagonal elements of the "hat" matrix: the name leverage, which is close in meaning to the word influence, was adopted to reflect this similarity. In the finite, sample approach to influence, the residual, or distance of the response to the fitted regression plane, is also relevant, so the name leverage seems misleading. One might prefer a more descriptive term, such as potential, which reflects the nonstochastic nature of these values.

The next principle is:

I4. Since the relevant quantities in influence analysis may be vector valued, a summarizing norm is usually required. The norm should (1) possess desirable statistical properties, such as invariance; (2) depend on the specific aspect of the analysis of interest and (3) the resulting values should be calibrated with respect to some external reference.

Cook's distance is an excellent example of an influence measure that satisfies this principle. In obvious notation, D_i can be written as

$$D_i = \frac{1}{p\hat{\sigma}^2}(\hat{\beta}_{(i)}-\hat{\beta})^T(X^TX)(\hat{\beta}_{(i)}-\hat{\beta}) = \frac{1}{p\hat{\sigma}^2}(\hat{Y}_{(i)}-\hat{Y})^T(\hat{Y}_{(i)}-\hat{Y})$$

This statistic is invariant under nonsingular linear transformations. Further, by its definition it measures either the change in the estimate of β, relative

to its variance, or the change in the fitted value vector. Finally D_i can be calibrated by comparison to confidence contours for $\hat{\beta}$. If $D_i \cong 1$, then deletion of the i-th case displaces the estimate of β to the edge of about a 50% confidence ellipsoid. Similarly, for $i \neq j$, D_i and D_j can be compared directly with each other, since both measures are made with respect to the same ellipsoids.

In contrast, the very similar statistic

$$\frac{1}{p}(\text{DFFITS}_i)^2 = \frac{1}{p\hat{\sigma}^2_{(i)}}(\hat{\beta}_{(i)}-\hat{\beta})^T(X^TX)(\hat{\beta}_{(i)}-\hat{\beta})$$

lacks the simple confidence contour interpretation so, in particular DFFITS_i and DFFITS_j cannot be compared directly (however, see Atkinson, 1981, for an alternative point of view). Similarly, the DFBETAS_{ij} are not invariant under nonsingular linear transformation, suggesting that they will be useful only in very special problems.

Many other norms for influence that satisfy I4 are given by Cook and Weisberg (1982, Section 3.5 and 5.2). These include essentially nonparametric norms, and a measure that uses the log likelihood function to define influence. This latter method is particularly useful in nonlinear problems where norms with elliptical contours are not easily justified.

The remarkable fact is that most sensible influence measures are very similar in practice (comparisons are given by Cook and Weisberg, Chapters 4 and 5). For the one at a time perturbation scheme, they all seem to depend on two basic building blocks, namely a Studentized residual and the diagonals of the hat matrix. Thus, three numbers per case (in Hocking's notation, h_i, either of t_i or t_i^*, and either of D_i or DFFITS_i) contain the relevant information concerning influence. The other measures that are popularly available are either practically equivalent to these, or else fail to satisfy the principles.

There is, or should be, an interplay between regression diagnostics and influence measures, since the influence measures are defined relative to a model, and diagnostics give information on the assumptions contained in the model. We could even dream of diagnostics for influence measures and vice versa. This process can, of course, be continued indefinitely. Rather than this, we should rely on appropriate plots to find such problems.

3. Concluding Comments

The number of techniques for regression diagnostics and for influence analysis is indeed very large. The diversity has three causes. First, the various concerns require the use of different statistics. This is especially true for the residuals, where several transformations of them are used. Second, several statistics that are practically identical, such as D_i, $DFFITS_i$ and the Johnson-Geisser measures will continue to compete because each has a different interpretation and will therefore have unequal appeal to different investigators. The analyst must try to understand the basis of these methods, and choose the one that seems most appropriate. Finally, there are other methods that simply are not as helpful as other methods because they fail to satisfy the principles. These are probably more successful at adding confusion than at adding information.

At Minnesota, we have taught the new diagnostic and influence methods since 1975, using locally written software. Our approach has always been to select the methods we think are the most useful; our selection of methods has changed over time. Students take up these methods well, and many of them become skillful and thoughtful data analysts who can use these methods in the ways intended: they are used as aids to understand a problem, and not as substitutes for independent thought.

Bibliography

AITCHISON, J. and SILVEY, S.D. (1960). "Maximum-likelihood estimation and associated tests of significance," Journal of the Royal Statistical Society, Series B, 22, 154-71.

ANDREWS, D.F. and PREGIBON, D. (1978). "Finding outliers that matter," Journal of the Royal Statistical Society, Series B, 40, 85-93.

ATKINSON, A.C. (1973). "Testing transformations to normality," Journal of the Royal Statistical Society, Series B, 35, 473-479.

_____(1981). "Robustness, transformations and two graphical displays for outlying and influential observations in regression," Biometrika, 68, 13-20.

_____(1982). "Regression diagnostics, transformations and constructed variables," (with discussion), Journal of the Royal Statistical Society, Series B, 44, 1-35.

BOX, G.E.P. (1980). "Sampling and Bayes' inference in scientific modelling and robustness," (with discussion), Journal of the Royal Statistical Society, Series A, 143, 383-430.

BOX, G.E.P. and COX, D.R. (1964). "An analysis of transformations," (with discussion), Journal of the Royal Statistical Society, Series A, 143, 383-430.

BOX, G.E.P. and TIDWELL, P.W. (1962). "Transformations of the independent variables," Technometrics, 4, 47-67.

BREUSCH, T.S. and PAGAN, A.R. (1979). "A simple test for heteroscedasticity and random coefficient variation," Econometrika, 47, 1287-1294.

COOK, R.D. and WEISBERG, S. (1982). Residuals and Influence in Regression. London and New York: Chapman-Hall.

_____ (1983). "Diagnostics for heteroscedasticity in regression," Biometrika, 70, (in press).

COX, D.R. (1977). "Nonlinear models, residuals and transformations," Math, Operationsforsch. Statist., Ser. Statistics, 8, 3-22.

DAVIES, R.B. and HUTTON, B. (1975). "The effects of errors in the independent variables in linear regression," Biometrika, 62, 383-91.

HOCKING, R.R. (1976). "The analysis and selection of variables in linear regression," Biometrics, 32, 1-40.

HODGES, S.D. and MOORE, P.G. (1972). "Data uncertainties and least squares regression," Applied Statistics, 21, 185-95.

HUBER, P. (1983). "Minimax aspects of bounded influence regression," Journal of the American Statistical Association, 78, (in press).

JARQUE, C.M. and BERA, A.K. (1980). "An efficient large sample test for normality of observations and regression residuals". Unpublished.

JOHNSON, W. and GEISSER, S. (1983). "A predictive view of the detection and characterization of influential observations in regression analysis," Journal of the American Statistical Association, 78, (in press).

MORAN, P.A.P. (1970). "On asymptotically optimum tests of composite hypothesis," Biometrika, 57, 47-55.

WEISBERG, S. (1980a). Applied Linear Regression. New York: Wiley.

_____ (1980b). "Comment on a paper by White and MacDonald," Journal of the American Statistical Association, 75, 28-31.